1 Introduction

In this Project we aim to explores how surface temperature changes relate to shifts in global and regional forest carbon stocks. Forests play really important role in global carbon cycle. we try to find that if there is any correlation between forest carbon stocks and Surface temperature changes.

2 Research Question

Is there a correlation between changes in forest carbon stocks and surface temperature changes over the decades globally or regionally?

3 Data Sources

3.1 Forest and Carbon Dataset

- **Metadata URL:** IMF Forest Data
- **Source:** International Monetary Fund (IMF)
- **Content:** *Country*, *Indicator*, *Unit*, annual data from 1992 to 2020.
- **Quality:** overall accurate, complete and consistent. this dataset allows for a inclusive analysis of how forest areas and carbon stocks have evolved over almost three decades.

3.2 Annual Surface Temperature Change Dataset

- **Metadata URL:** IMF Surface Temperature Data
- **Source:** International Monetary Fund (IMF)
- **Content:** *Country*, *Indicator*, *Unit*, annual data from 1961 to 2022.
- **Quality:** this dataset provides a long term perspective on temperature variations, important for analyzing climate trends.

3.3 Data Structure and Quality

- **Format:** CSV files and Stored in an SQLite database.
- **Tables:**
  - *Annual_Surface_Temperature_Change:* Columns for *Country*, *Indicator*, and annual data from 1961 to 2022.
  - *Forest_and_Carbon:* Columns for *Country*, *Indicator*, and annual data from 1992 to 2020.
  - *Temp_Change_Diff:* Annual differences in temperature change for each country.
  - *Carbon_Stocks_Diff:* Annual differences in carbon stocks for each country.
- **Quality:** Mostly complete, consistent, and accurate. The dataset ensures reliability in subsequent analyses.

3.4 Licenses

datasets can be used and distributed with proper citation. IMF Data Terms of Use. This open data policy facilitates academic and public research.

4 Data Pipeline

4.1 Description

The pipeline fetches, cleans, transforms, and stores data in a SQLite database using Python, pandas, and SQLite making data ready for analysis.

1. **Download Data:** Downloaded and Saved the data locally
2. **Data Cleaning:** clean out the data by removing the rows with missing values, convert data to numeric format.
3. **Data Transformation:** calculated annual differences in temperature and carbon stocks.
4. **Data Storage:** finally after transforming Stored the data in an SQLite database for analysis.

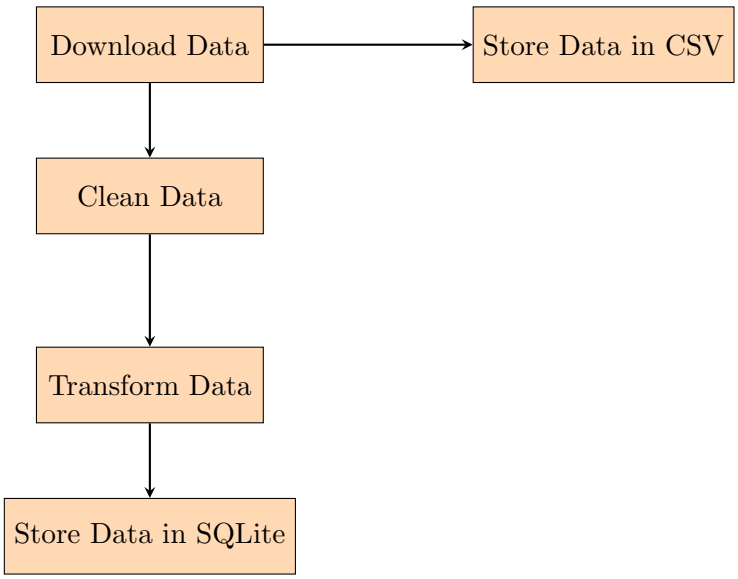

Figure 1: Data Pipeline Diagram

4.2 Transformation and Cleaning Steps

Download and Read CSVs: Data is downloaded and read into pandas DataFrames.

- **Removed Unnecessary Columns:** Simplified the datasets by removing columns that are not important.
- **Dropped Missing Values:** set the seal on data integrity by removing records with missing values.
- **Converted Data Types:** changed required columns type to numeric format for more better analysis.
- **Set Index:** setted the *Country* column as the index.
- **Calculated Annual Differences:** Computed yearly changes to observe trends over time.
- **Stored in SQLite Database:** Saved the cleaned data in SQLite for better storage and querying.

4.3 Problems Encountered

- **Missing Values:** Rows with NaNs were removed to maintain data quality.
- **Data Type Inconsistencies:** Ensured numerical columns were correctly formatted for accurate analysis.

5 Result and Limitations

5.1 Output Data

The pipeline output data has:

- Annual temperature change data for several countries.
- Forest and carbon stock data for several years.
- Annual differences in temperature changes and carbon stocks, for detailed trend analysis.

5.2 Data Format

Reason for Choosing SQLite:

- **Efficiency:** Efficient storage and querying.
- **Portability:** Easily transferable and shareable.
- **Ease of Use:** Simple to use with with pandas for data manipulation.

5.3 Critical Reflection and Potential Issues

- **Strengths:** Clean and consistent data, efficient querying capabilities.
- **Potential Issues:**
  - **Data Completeness:** By Dropping rows with missing values may lead to loss of valuable information.
  - **Linear Assumptions:** Annual differences assume linear changes, which sometimes overlook complex patterns.
  - **Geographical and Temporal Coverage:** data is limited to certain regions or periods which may affect trend analysis.
  - **Analysis Limitations:** The linear assumption in annual differences might oversimplify real world dynamics.
- **Future Adaptations:** The pipeline might need modifications to coupe up with changes in data sources or to integrate additional datasets for a more comprehensive analysis.